Supporting Information

Sims et al. 10.1073/pnas.0909377106

SI Text

Supplementary Materials and Methods

FFP Alignment-Free Genome Comparison. All of the genomic partitions were compared with one another by using the FFP alignment-free method. The genome sequence is converted first to an purine–pyrimidine base alphabet. This strategy considers only transversion events between the two types of bases in the genome (purines and pyrimidines). This is based on the observation that mutational rates vary broadly across species, but the transversion rate is much slower than the transition rate (1). Furthermore, this strategy dramatically reduces the computational memory requirement that otherwise prevents application of the FFP method for the large genomes of mammals. The sequence is divided into overlapping features, or *l*-mers, of a given length or resolution, *l*. The feature frequencies are then used as the basis for genome comparison. To count the frequencies of each feature in the genome, a sliding window of length *l* is run through the sequence from position 1 to $n - l + 1$. When counting, *l*-mers continue over the whole genome, but the sliding window is not allowed to span over gaps or "x" characters from sequence concatenation between contigs. Also, the forward and reverse complement word (features) are considered equivalent. The counts are tabulated in the vector $\mathbf{C}_l$ for all possible features of length *l*,

$$C_l = \langle c_{l,1}, \ldots, c_{l,K} \rangle \qquad [1]$$

where *K* is the number of all possible features of length *l*. For an odd length *l*, the number of features is

$$K = 2^{l-1} \qquad [2]$$

and for even length *l*

$$K = (2^l + 2^{l/2})/2 \qquad [3]$$

Note that Eqs. **S2** and **S3** are a result of forward/reverse complement equivalency. The raw frequency counts are normalized to form a probability distribution vector or FFP, $\mathbf{F}_l$, giving the relative abundance of each *l*-mer. This normalization removes small genome length differences as a factor in the comparison. The distance between two probability vectors $\mathbf{P}_l$ and $\mathbf{Q}_l$ is calculated by using the Jensen-Shannon (JS) divergence (2),

$$JS_l(\mathbf{P}_l, \mathbf{Q}_l) = \frac{1}{2} KL(\mathbf{P}_l, \mathbf{M}_l) + \frac{1}{2} KL(\mathbf{Q}_l, \mathbf{M}_l) \qquad [4]$$

where $\mathbf{M}_l = (\mathbf{P}_l + \mathbf{Q}_l)/2$ and KL is the Kullback–Leibler divergence (3),

$$KL(\mathbf{P}_l, \mathbf{M}_l) = \sum_{i=1}^{K} p_{l,i} \log_2 \frac{p_{l,i}}{m_{l,i}} \qquad [5]$$

A matrix composed of pair-wise divergences is used as input to the neighbor-joining tree-construction method.

Feature Filtering for Mammalian Genomes. Two forms of feature filtering were applied for each class of genome partitions: (*i*) feature complexity and (*ii*) feature frequency. Mammal genomes contain a large fraction of sequence which is repetitive or of low complexity. The complexity of a feature, K_f, is determined by comparing its size in bytes, before and after Limpel–Ziv lossless compression (4).

$$K_f = |s - s_{compress}| \qquad [6]$$

The compression is implemented by using the gzip utility (gzip −9). The complexity of *l*-mers for a given *l* is normally distributed, and we choose only the high-complexity features, where K_f is greater than the one standard deviation (σ) below the average complexity (μ), or $\mu - \sigma$. Also, high-frequency features should be disregarded because they are not sensitive for distinguishing different genomes, and these features dominate the JS divergence score. The average and standard deviation of the count values, $c_{l,i}$ for all genomes were calculated for each class of genome partitions, and we chose only those features with $c_{l,i} < \mu + \sigma$.

Assessment of Feature Correlation. Many features in each FFP are correlated, that is to say, when frequencies are examined across species the correlation among features is high. This correlation arises from two sources: (*i*) from the "sliding frame" method of identifying features, where features may be a connected part of a larger motif, and (*ii*) from possible redundancy related to an evolutionary signal. In the latter case, feature frequencies are not independent of each other, but are dependently related to each other via common lines of species divergence. We wanted to assess the extent of correlation to determine whether removal of a large number of features through the process of filtering (≈30–40% for each class of genome partitions) could be eliminating the phylogenic signal. For a particular *l*, two features were determined to be redundant at two levels if they have $\rho = 1$ or $\rho > 0.98$ (Spearman's rank correlation) in frequencies across species. The former threshold is equivalent to equal rankings and the latter to a difference in two ranks. Calculating a large feature-correlation matrix is prohibitive, so redundancy was estimated by repeatedly (*n* times) sampling a smaller set of *m* features and calculating this smaller correlation matrix. If the occurrence of redundancy in the sampled matrix is thought of as a sample from a Poisson process, the expected total percent redundancy of the complete matrix can be estimated via maximum likelihood :

$$R = \frac{100\,{}_kC_2}{kn\,{}_mC_2} \sum_{i=1}^{n} r_i \qquad [7]$$

where r_i is the number of feature pairs that are correlated above the threshold in sample *i*, *k* is the total number of features (Eqs. **S2** and **S3**), and $n = 10^4$, and $m = 10^3$. This procedure was repeated for all classes of genome partitions. The percent redundancies for each of the partitions for $l = 18$ at $\rho = 1$ were: whole, 4.15%; intronic, 5.04%; nongenic, 4.24%; exonic, 4.38%. At $\rho > 0.98$ the percent redundancies were: whole, 43.2%; intronic, 48.1%; nongenic, 41.3% .

1. Collins DW, Jukes TH (1994) Rate of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* 20:386–396.
2. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Info Theory* 37:145–151.
3. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22:79–86.
4. Ziv J, Lempel A (1978) Compression of individual sequences via variable-rate coding. *IEEE Trans Info Theory* 24:530–536.

Table S1. Gene-tree topology types

K2-Neighbor-joining tree types			
Type I		Type II	
Chrm1	cholinergic receptor, muscarinic 1	Chrm1-*RY*	cholinergic receptor, muscarinic 1
Clock-*RY*	circadian locomoter output cycles kaput	Clock	circadian locomoter output cycles kaput
Cnr1-*RY*	cannabinoid receptor 1	Cnr1	cannabinoid receptor 1
Lamc2-*RY*	laminin, gamma 2	Lamc2	laminin, gamma 2
Nd1	MT-NADH dehydrogenase 1	Nd1-*RY*	MT-NADH dehydrogenase 1
Plcb4-*RY*	phospholipase C, beta 4	Plcb4	phospholipase C, beta 4
Rbp3	retinol binding protein 3, interstitial	Rbp3-*RY*	retinol binding protein 3, interstitial
Zfx-*RY*	zinc finger protein, X-linked	zfx	zinc finger protein, X-linked
Cenpb-*RY*	centromere protein B	Cenpb	centromere protein B
Prom1	prominin 1	Ptprb	protein tyrosine phosphatase, receptor type, B
Bmi1	Bmi1 polycomb ring finger oncogene	Cytb	MT-cytochrome b
Ets1	v-ets erythroblastosis virus E26 oncogene homolog 1	Eftud2	elongation factor Tu GTP binding domain containing 2
Ets2	v-ets erythroblastosis virus E26 oncogene homolog 2	Dscam	Down’s syndrome cell-adhesion molecule
Rag1	recombination activating gene 1	Runx1	runt-related transcription factor 1
Rag2	recombination activating gene 2	Tyr	tyrosinase
Bdnf	brain-derived neurotrophic factor	Pax6	paired box 6
Eftud1	elongation factor Tu GTP binding domain containing 1	Fgg	fibrinogen gamma chain
gzmb	granzyme B precursor	Adora3	adenosine A3 receptor
Csnk2b	casein kinase 2B	Adra2b	adrenergic receptor, alpha 2b
Kcnj5	potassium inwardly-rectifying channel, J5	Akirin2	akirin 2
Atxn1-*RY*	ataxin 1	App	amyloid beta (A4) precursor protein

RY indicates the topology of the tree after the alignment was reduced to purine/pyrmidine characters.